© 1994 Oxford University Press

Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes

J.R.Lobry* and C.Gautier

Laboratoire de Biométrie, CNRS URA 243, Université Claude Bernard, 43 Bld. du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France

Received March 28, 1994; Revised and Accepted July 11, 1994

ABSTRACT

Multivariate analysis of the amino-acid compositions of 999 chromosome-encoded proteins from *Escherichia coli* showed that three main factors influence the variability of amino-acid composition. The first factor was correlated with the global hydrophobicity of proteins, and it discriminated integral membrane proteins from the others. The second factor was correlated with gene expressivity, showing a bias in highly expressed genes towards amino-acids having abundant major tRNAs. Just as highly expressed genes have reduced codon diversity in protein coding sequences, so do they have a reduced diversity of amino-acid choice. This showed that translational constraints are important enough to affect the global amino-acid composition of proteins. The third factor was correlated with the aromaticity of proteins, showing that aromatic amino-acid content is highly variable.

INTRODUCTION

This paper investigates the amino-acid usage in *Escherichia coli* proteins, to describe general trends and their biological implications. The method used, correspondence analysis, has also been used to analyze codon usage by Grantham and colleagues (1–3, review in 4). The first factor underlying variations in codon usage is the genome of origin. In addition, there is a considerable within-species codon usage variability. Among *E.coli* genes this diversity is linked to gene expressivity: genes with a potentially high expression level are biased towards the subset of codons that are best recognised by the most abundant tRNA species (5). In contrast with codon usage, the interspecific variability in amino-acid usage is low (3). The present study focuses on amino-acid usage of proteins from a single species, *E.coli*, because a large body of sequence data is available for this species.

MATERIALS AND METHODS

Data set

The data set was 999 protein sequences encoded by genes on the *E.coli* chromosome, corresponding to a total of 385,404 amino-acids. As this is about 25% of the estimated total number of chromosome-encoded proteins, the sample is large enough to be representative. The nonoverlapping ECOSEQ6 collection (6) was structured (7) using the entity-relationship model of ACNUC (8–10). The retrieval system, Query, associated with ACNUC, allows elaborate sequence managements. The ECOSEQ6 collection contains the sequences of a single allele per locus, so that there is no overweighting due to sequence redundancy or DNA polymorphism. This is not a negligible problem since, for instance, there are 16 complete sequences of the *gnd* locus of *E.coli* in GenBank (11) release 78. The disadvantage is that the allele sequences in Rudd's collection are from different strains, leaving open the possibility of intraspecific variations affecting results. There are not yet enough data to answer this question, but there seems to be very little polymorphism at the amino-acid level, about 1% for the average number of amino-acid differences per site between two alleles (12).

Plasmid-encoded proteins are not included in Rudd's collection. This minimizes the horizontal gene transfer effect, which is more likely for plasmid-encoded protein. The amino-acid usage of proteins encoded by genes recently incorporated in the *E.coli* genome may differ from native *E.coli* proteins.

Partial sequences (7%) were discarded because the amino-acid composition of a fragment could be atypical of the whole protein composition. Poorly documented open reading frames (12%) were discarded to help analysis of results. The Rudd nomenclature, by which most unidentified ORFs are given a name starting with 'y', ensure their easy removal. Information on the remaining sequences is, however, highly variable. Proteins with fewer than 100 amino-acids (5%) were excluded to minimize influence of stochastic variations in the amino-acid compositions of small peptides. The threshold value of 100 amino-acids is roughly the minimum size for a protein to have an enzymatic function (13).

The N-terminal methionine was not removed. This is an arbitrary choice because the rules that govern the removal of N-formylmethionine are not completely understood (14). This choice did not noticeablely alter the results, there were negligible variations only for small proteins with a low methionine frequency. The special case of selenocystein was not handled

*To whom correspondence should be addressed because it is too rare; there are only three known selenopolypeptides in *E.coli* (15). Lastly, post-translational modifications were not taken into account.

Multivariate analysis

The χ^2 metric was used as a measure of the distance between the amino-acid composition of two proteins. Correspondence analysis can then extract orthogonal linear combinations of amino-acid frequencies that best summarize the data. These trends are optimal because they take into account most of the initial variability (16, 17). The squared distance between two sequences x and y is defined as:

$$d^2(x,y) = n_{..} \sum_{i=1}^{20} \frac{1}{n_{\cdot i}} \left(\frac{n_{xi}}{n_{x\cdot}} - \frac{n_{yi}}{n_{y\cdot}}\right)^2$$

where n_{xi} and n_{yi} are the number of amino-acids of kind i in sequence x and y, $n_{x.}$and $n_{y.}$ are the total number of amino-acids in sequences x and y, $n_{.i}$ the total number of amino-acids of kind i in the dataset and $n_{..}$ the total number of amino-acids in the dataset. The advantage of the χ^2 metric over the usual Euclidian distance used in principal component analysis of compositional data (18), is that information on rare amino-acids are not masked by frequent amino acids because of the $1/n_{.i}$ weighting.

The correspondence analysis was computed with the program MacMul (19, 20) on a Macintosh plus. The results were checked by running a different program (21) on a different computer (Sun SPARCStation 10) to ensure that there were no computational errors. Analysis of results was facilitated by the interactive DIGIT software (22). The absence of bias due to the low frequencies of rare amino-acids was checked by removing them and repeating the analysis.

Identification of protein characters

Three scores were computed for each protein to help interpret the results. The GRAVY score (23) is an estimate of the overall hydrophobicity of the protein, the highest scores indicating a hydrophobic character. The GRAVY score is a linear combination of amino-acid relative frequencies:

$$\text{GRAVY} = \sum_{i=1}^{20} \alpha_i f_i$$

where f_i is the relative frequency of amino-acid of kind i in the protein and α_i the hydropathy index of this amino-acid (23).

The codon adaptation index (CAI) is an empirical measure of synonymous codon usage bias (24), which is positively correlated with the expressivity level of genes.

$$\ln(\text{CAI}) = \sum_{i=1}^{61} f_i \ln w_i$$

where f_i is the relative frequency of codon of kind i in the coding sequence, and w_i the ratio of the frequency of codon of kind i to the frequency of the major codon for the same amino-acid, as estimated from examining 25 highly expressed genes (24). Here, CAI has the advantage over other indices, such as the Mean Number of tRNA Discrimination per elongation cycle (5), of being almost independent of amino-acid frequencies.

The AROMATICITY is the relative frequency of aromatic amino-acids,

$$\text{AROMATICITY} = \sum_{i=1}^{20} \delta_i f_i ,$$

where f_i is the relative frequency of amino-acid of kind i in the protein and $\delta_i = 1$ when the amino-acid is aromatic (Phe, Tyr, Trp) and $\delta_i = 0$ otherwise.

RESULTS

Glbal amino-acid composition

The mean amino-acid composition of the proteins in the dataset (Table 1) was found to be very similar (r = 0.95) to that reported previously (25). The results are also consistent (r = 0.89) with the experimentally determined composition of the total proteins

Table 1. Average amino-acid composition (% ± SD)

AA	Total	IMP	non IMP	N	C
Ala	9.7 ±2.6	10.4 ±2.3	9.6 ±2.7	8.8±3.5	9.6
Arg	5.8 ±2.2	3.7 ±1.3	6.0 ±2.2	4.4±2.1	5.5
Asn	3.8 ±1.4	3.0 ±1.2	3.9 ±1.4	10.5±3.0	9.0
Asp	5.3 ±1.8	2.5 ±0.8	5.7 ±1.5		
Cys	1.2 ±1.0	1.0 ±0.7	1.2 ±1.0	1.4±1.1	1.7
Gln	4.3 ±1.8	2.6 ±1.2	4.5 ±1.7	10.6±3.5	9.8
Glu	6.1 ±2.3	2.5 ±1.1	6.6 ±2.0		
Gly	7.5 ±2.1	8.7 ±2.0	7.3 ±2.0	8.1±2.8	11.5
His	2.3 ±1.2	1.6 ±1.0	2.4 ±1.2	2.1±1.2	1.8
Ile	5.9 ±1.9	8.1 ±2.0	5.7 ±1.7	5.0±2.0	5.4
Leu	10.2 ±2.7	13.2 ±2.9	9.9 ±2.5	8.1±2.8	8.4
Lys	4.7 ±2.3	2.9 ±1.2	4.9 ±2.0	6.5±3.1	6.4
Met	2.8 ±1.2	3.9 ±1.3	2.7 ±1.1	1.9±1.1	2.9
Phe	3.8 ±1.6	6.3 ±1.9	3.5 ±1.3	3.8±1.6	3.5
Pro	4.4 ±1.6	4.0 ±1.5	4.4 ±1.6	4.7±1.9	4.1
Ser	5.5 ±1.7	6.2 ±1.7	5.5 ±1.7	6.8±2.7	4.0
Thr	5.3 ±1.5	5.3 ±1.4	5.3 ±1.5	5.9±2.2	4.7
Trp	1.3 ±1.0	2.6 ±1.2	1.2 ±0.9	1.1±0.8	1.1
Tyr	2.7 ±1.3	3.0 ±1.3	2.7 ±1.3	3.3±1.5	2.6
Val	7.3 ±1.9	8.6 ±1.9	7.1 ±1.8	7.0±2.2	7.9

IMP is the group of 114 integral membrane proteins given in table 3. Column N contains the results previously reported (25) and column C is the experimentally determined total protein composition (26). Since determination of protein composition requires the hydrolysis of all amide bonds, the relative amounts of Asp:Asn and Glu:Gln cannot be estimated, and their values are usually assumed to be 1:1. Here, the ratios were found to be about 3:2, showing that the acidic form was more abundant.

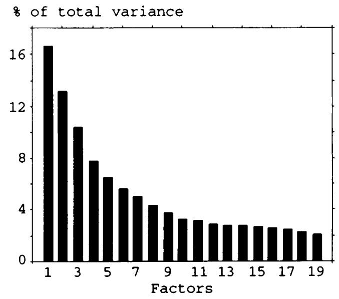

Figure 1. Factors of the correspondence analysis ranked in decreasing order of the fraction of total variance they accounted for.

of *E.coli* (26), although these results are not directly comparable, because of inequal protein concentrations *in vivo*.

On the basis of their average frequencies, amino-acids can be classified as very rare (Trp, Cys), rare (Tyr, Met, His), frequent (Gly, Val), very frequent (Leu, Ala) and intermediary for the remainder. In general, aliphatic amino-acids occur frequently, while aromatic or sulfur containing amino-acids are rare.

The relative frequencies of amino-acids within protein have unimodal, nearly symetric distributions, except for rare amino-acids (Trp and Cys), because they are quite often absent from a protein (about 10% of proteins lacked Trp or Cys in our data set).

Selection of factors

The relative importance of factors, as juged by the difference with their following factor, was found to vanish with factor 4 (Figure 1). The three first factors, which accounted for 40% of the total variability of amino-acid composition of *E.coli* proteins, were then further analysed. These factors are defined in Table 2.

Table 2. Definition of the first three factors of the correspondence analysis (F1, F2 and F3)

AA	F1	F2	F3	AF	RF	F1.RF
Ala	-0.319	-0.306	-1.515	35	0.114	-0.036
Arg	+1.395	+1.538	-0.039	5	0.016	+0.023
Asn	+0.239	-0.964	+1.425	14	0.046	+0.011
Asp	+1.428	-0.680	+0.696	14	0.046	+0.065
Cys	+0.102	+1.759	+1.036	1	0.003	+0.000
Gln	+1.204	+1.316	-0.441	16	0.052	+0.063
Glu	+1.902	-0.314	-0.475	9	0.029	+0.056
Gly	-0.740	-0.679	-0.117	20	0.065	-0.048
His	+0.949	+1.496	+0.875	1	0.003	+0.003
Ile	-1.019	-0.853	-0.333	10	0.033	-0.033
Leu	-0.582	+1.171	-0.764	33	0.108	-0.063
Lys	+1.026	-2.302	+0.444	19	0.062	+0.064
Met	-1.107	-0.350	-0.497	6	0.020	-0.022
Phe	-1.888	-0.037	+1.242	8	0.026	-0.049
Pro	+0.123	+0.727	+0.399	25	0.082	+0.010
Ser	-0.527	+0.321	+0.248	29	0.095	-0.050
Thr	-0.097	-0.255	+0.372	27	0.088	-0.009
Trp	-2.233	+2.668	+2.687	1	0.003	-0.007
Tyr	-0.121	-0.347	+3.311	7	0.023	-0.003
Val	-0.520	-0.613	-0.825	26	0.085	-0.044
Σ				306	1.000	$\boxed{-0.069}$

The computation of F1 score for MalM (accession number = X04477) is explained. AF is the absolute frequency of amino acids in MalM, including initial methionine, and RF the relative frequency. The score of MalM on the first factor (−0.069) is found by summing the products of F1 by RF.

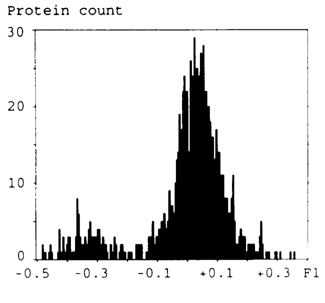

Figure 2. Distribution of scores for correspondence analysis factor 1. The minor peak (11% of total) contains integral membrane proteins.

Table 3. List of proteins (score < −0.2) in the minor peak of F1 scores

Prot.	F1	FUNCTION
CyoD	-0.483	Component of the cytochrome o ubiquinol oxidase
SdhD	-0.482	Anchor polypeptide of succinate dehydrogenase
MvrC	-0.473	Methyl viologen resistance
TnaB	-0.460	Transport of Tryptophan
BicA	-0.456	Involved in bicyclomycin resistance
LacY	-0.455	Transport of lactose
DmsC	-0.450	Anchor polypeptide of the anarobic dimethylsulfoxide reductase
NirC	-0.431	Transport of nitrite
CodB	-0.427	Transport of cytosine
RhaT	-0.427	Transport of L-rhamnose
NupG	-0.426	Transport of nucleosides
MreD	-0.426	Involved in the formation of rod shape of the cell
AraJ	-0.422	Transport of arabinose polymers (putative)
AppB	-0.415	Component of a third cytochrome oxidase (putative)
Mtr	-0.413	Transport of tryptophan
FrdD	-0.412	Anchor protein of the fumarate reductase complex
MglC	-0.409	Transport of beta-methylgalactoside
PstC	-0.404	Transport of phosphate
NarK	-0.402	Transport of nitrate
TrkG	-0.398	Transport of potassium
CadB	-0.396	Transport of lysine and cadaverine (putative)
CvpA	-0.396	?
CyoB	-0.395	Component I of the cytochrome o ubiquinol-8 oxidase
AraH	-0.392	Transport of L-arabinose
PotE	-0.391	Transport of putrescine
CyoC	-0.390	Component III of the cytochrome o ubiquinol-8 oxidase
LivH	-0.382	Transport of branched-chain amino acids
UgpE	-0.377	Transport of sn-glycerol-3-phosphate
Rfe	-0.374	Synthesis of lipid I
CyoE	-0.372	Component of the cytochrome o ubiquinol-8 oxidase
PutP	-0.370	Transport of proline
CdsA	-0.370	Synthesis of polar head of phospholipids
CynX	-0.369	?
UdpA	-0.368	Transport of sn-glycerol-3-phosphate
TrkH	-0.367	Transport of potassium
NarV	-0.366	Component of the second nitrate reductase
PotC	-0.365	Transport of spermidine and putrescine
PheP	-0.365	Transport of phenylalanine
LysP	-0.365	Transport of lysine
TdcC	-0.364	?
UhpC	-0.363	Transport of hexoses phosphates
HycC	-0.362	Component of the formate hydrogene lyase
FhuB	-0.362	Transport of ferric hydroxamate
TyrP	-0.361	Transport of tyrosine
LspA	-0.361	Lipoprotein signal peptidase
AroP	-0.360	Transport of aromatic amino-acids
GabP	-0.360	Transport of 4-aminobutyrate
PotB	-0.358	Transport of spermidine and putrescine
GltS	-0.357	Transport of glutamate
CydB	-0.353	Component II of cytochrome d
MraY	-0.352	?
BtuC	-0.348	Transport of vitamin B12
NhaA	-0.347	Transport of sodium
KdpA	-0.346	Transport of potassium
MalG	-0.339	Transport of maltose
SdhC	-0.339	Anchor polypeptide of succinate dehydrogenase
RbsC	-0.339	Transport of ribose
EmrB	-0.337	Involved in multidrug resistance
FucP	-0.337	Transport of L-fucose
MelB	-0.330	Transport of melibiose
FepG	-0.328	Transport of ferric enterobactin
ManY	-0.328	Transport of manose
GlpT	-0.327	Transport of glycerol-3-phosphate
HyaC	-0.326	Component of hydrogenase 1
MrdB	-0.325	Penicillin binding protein 5
CelB	-0.324	Transport of beta-glucoside sugars
GlpF	-0.324	Transport of glycerol
CysU	-0.324	Transport of sulfate
NarI	-0.322	Anchor polypeptide for cytochrome bNR
FecD	-0.321	Transport of ferric dicitrate
XylE	-0.321	Transport of Xylose
PstA	-0.318	Transport of phosphate
LivM	-0.317	Transport of branched-chain amino-acids
UhpT	-0.313	Transport of hexose 6-phosphate
HycD	-0.313	Component of hydrogenase 3
FrdC	-0.310	Anchor protein of fumarate reductase
UbiA	-0.309	Synthesis of ubiquinone
KgtP	-0.307	Transport of alpha-ketoglutarate
MalX	-0.306	?
PanF	-0.304	Transport of pantothenate
FtsW	-0.304	Involved in cell disvision
GltP	-0.300	Transport of glutamate
FepD	-0.298	Transport of ferric enterobactin
ProW	-0.298	Transport of glycine-betaine and proline
PtsG	-0.297	Transport of glucose
FecC	-0.295	Transport of ferric dicitrate
AraE	-0.294	Transport of L-arabinose
RfaL	-0.294	?
SbmA	-0.294	Sensitivity to microcin B17
PgsA	-0.291	Synthesis of phospholipids
FdnI	-0.288	Component of nitrate inductible formate dehydrogenase
DgkA	-0.283	Diacylglycerol kinase
CybB	-0.282	Cytochrome b561
AscF	-0.282	Transport of beta-glucoside sugars
DedA	-0.280	?
SecY	-0.277	Involved in protein export
PhnE	-0.276	Transport of phosphate
GlpG	-0.273	Component of the aerobic glycerol-3-phosphate dehydrogenase
NhaB	-0.272	Transport of sodium
BglF	-0.269	Transport of beta-glucoside sugars
DsbB	-0.254	?
CysW	-0.254	Transport of sulfate and thiosulfate
AppC	-0.253	Component of the third cytochrome oxidase (putative)
BetT	-0.244	Transport of choline
NagE	-0.242	Transport of N-acetylglucosamine
PntB	-0.241	Component of the pyridine nucleotide transhydrogenase
CydA	-0.240	Component I of cytochrome d
MalF	-0.239	Transport of maltose
SrlA	-0.225	Transport of D-glucitol
CutE	-0.224	Transport of copper
ManZ	-0.220	Transport of manose
PhnQ	-0.219	Transport of phosphate
FimH	-0.217	Regulation of the lenght and number of typel fimbriae
FruA	-0.210	Transport of fructose

These integral membrane proteins are involved in transport, anchoring of dehydrogenases, and synthesis of lipid bilayer components.

Factor 1 (F1)

The first, and thus most important, factor of the correspondence analysis accounted for 17% of the total variability of amino-acid composition of *E.coli* proteins. The protein F1 scores had a bimodal distribution (Figure 2), indicating that the amino-acid frequencies in the dataset were heterogeneous. The minor peak (11%) contained only integral membrane proteins (Table 3).

Factor 1 was highly correlated (r = 0.90, $p < 10^{-4}$) with the GRAVY score (figure 3). Direct comparison of the GRAVY score and the F1 score coefficients (figure 4) showed a major difference only for Trp. Another difference is that the GRAVY scale assigns the same value to Glu, Gln, Asp and Asn. The coefficients for Glu, Gln, Asp were found to be quite similar in the F1 score, but the coefficient for Asn was different.

Factor 2 (F2)

The second factor accounted for 13% of the variability in amino-acid compositions. Protein scores on this second factor had a

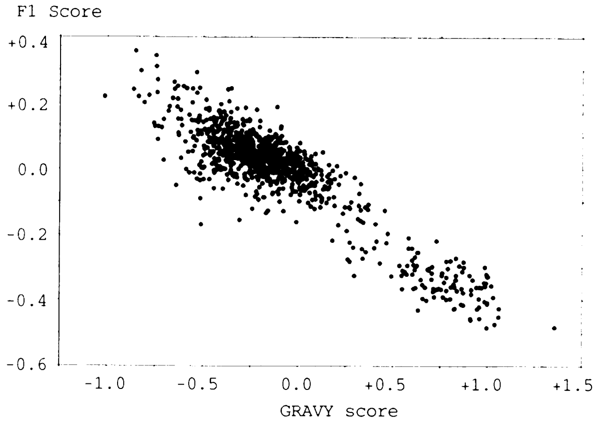

Figure 3. Correlation of the global hydrophobicity of proteins (GRAVY score) with the correspondence analysis factor 1. Each point represents a protein, the bottom right group is the integral membrane protein group.

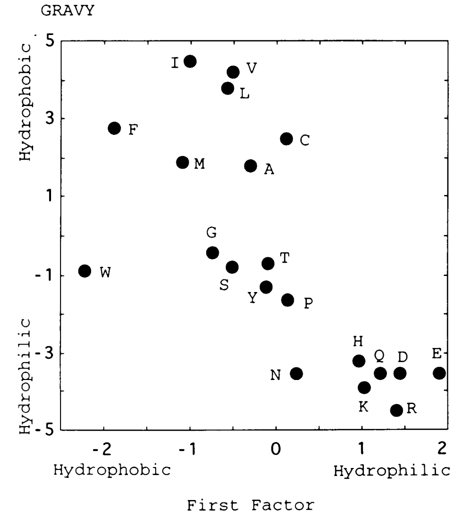

Figure 4. The coefficients for the GRAVY score and for the correspondence analysis factor 1, for the 20 amino-acids.

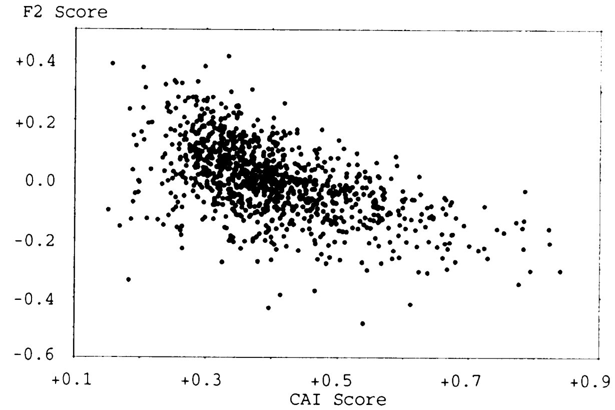

Figure 5. Correlation of the codon adaptation index (CAI) with the correspondence analysis factor 2. Each point represents a protein. Highly expressed genes have a high CAI value.

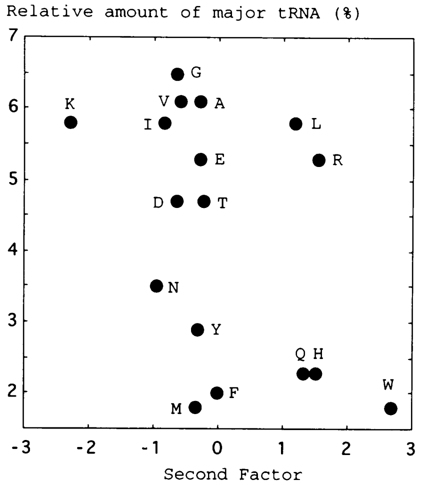

Figure 6. The intracellular concentrations of the major tRNA of amino-acids (36) and the coefficient for correspondence analysis factor 2. The concentrations of the major tRNA for Ser, Pro and Cys were not determined.

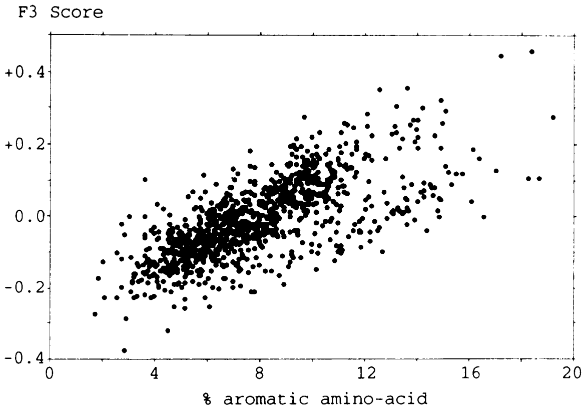

Figure 7. Correlation of the aromaticity with the correspondence analysis factor 3. Each point represents a protein.

unimodal, nearly symmetrical distribution. The F2 scores were correlated (r = 0.55, $p < 10^{-4}$) with the CAI scores (Figure 5). The general trend was that proteins with a high CAI value had low F2 scores. This result is highly surprising as CAI score is almost independent of amino-acid composition of the protein: CAI measures the codon usage bias cumulated for each amino acid. Hence, amino acid composition correlates with the choice of codon among synonymous sets.

A comparison of amino-acid F2 coefficients and major tRNAs concentrations (Figure 6) showed three notable exceptions. Lys was more enriched than expected from the relative frequency of its major tRNA, Leu and Arg were avoided despite the relative abundance of their major tRNA.

Factor 3 (F3)

The third factor accounted for 10% of the variability in amino-acid compositions. Protein scores on this second factor had a unimodal, nearly symmetrical distribution. The F3 scores were correlated (r = 0.70, $p < 10^{-4}$) with the aromaticity scores (Figure 7). The general trend was that proteins enriched in aromatic amino-acids had high F3 scores.

DISCUSSION

The pattern of amino-acid usage was very different from the pattern of codon usage. Analysis of the codons in the coding sequences of *E.coli* emphasises the contrast between lowly and highly expressed genes, with the optimal codons in highly expressed genes. But, as the table of amino-acid frequencies is obtained directly from the table of codon frequencies by summing columns, it seems surprising that the factors reported here have not been described before. One reason is that the column summing which transforms the codon frequency table into the amino-acid frequency table is very special in that frequent codons are summed with rare codons. As the contrast between rare and frequent codons is very important, the amino-acid tendencies are hidden in the least important factors of the codon multivariate analysis.

Integral membrane proteins are known to be enriched in hydrophobic amino-acids. Our correspondence analysis confirmed this and showed that this is the most important factor underlying variations in the global amino-acid composition of *E.coli* proteins.

As factor 1 clearly discriminates integral membrane proteins from the others, computing its value for a new open reading frame could indicate if it codes for an integral membrane protein (a complete example of the computation is given in Table 2). For instance, the protein CutE involved in copper transport in *E.coli* has an F1 score of −0.22. This suggests that it is an integral membrane protein, and not an intracellular protein (27). This prediction of an integral membrane protein is expected to occur for 1/10$^{\text{th}}$ of the *E.coli* coding sequences. Peripheral membrane proteins cannot be identified on the basis of their average amino-acid frequencies because the contribution of membrane-spanning segments to the overall amino-acid composition of the protein is not always sufficient (28).

Factor 2 showed that there was a bias in amino-acid usage for highly expressed genes. There is experimental evidence that the total amount of tRNA for a particular amino-acid parallels the total usage of that amino-acid in proteins for *E.coli* and *Mycoplasma capricolum* (29). Our results also show that proteins encoded by highly expressed genes tend to use amino-acids whose major tRNA are abundant. This bias is not negligible, since it is the second factor accounting for variability of the amino-acid variability of *E.coli* proteins. This bias was previously observed in studies on much smaller samples of *E.coli* proteins (30−32).

Table 4. Last 10% of CAI distribution

Prot.	CAI	FUNCTION
GapA	0.840	glyceraldehyde-3-phosphate dehydrogenase (glycolysis)
OmpC	0.824	major outer membrane protein (porin)
TufA	0.822	elongation factor EF-TU (translation)
MopA	0.794	refolding of protein under stress condition (translation related)
RpsI	0.785	ribosomal protein S9
Pfl	0.784	pyruvate formate-lyase (nonoxidative conversion of glucose)
RpsA	0.784	ribosomal protein S1
RpsB	0.782	ribosomal protein S2
Tsf	0.777	elongation factors EF-Ts (translation)
OmpA	0.772	major outer membrane protein (porin)
FusA	0.753	elongation factor EF-G (translation)
TpiA	0.743	triosephosphate isomerase (glycolysis)
RplI	0.729	ribosomal protein L9
SodA	0.724	manganese superoxide dismutase (radicals destruction)
DnaK	0.723	Major heat shock protein (DNA replication)
Tig	0.715	chaperone (protein export)
PykF	0.701	pyruvate kinase I (glycolysis)
Pal	0.692	peptidoglycan-associated lipoprotein (structure)
Pnp	0.680	polynucleotide phosphorylase: mRNA degradation (transcription)
PpiB	0.679	peptidyl-prolyl cis-trans isomerase (protein folding)
RplT	0.675	ribosomal protein L20
GlyA	0.674	serine hydroxymethyltransferase (purines & lipids synthesis)
RplM	0.674	ribosomal protein L13
AceE	0.673	pyruvate dehydrogenase (glycolysis)
AtpA	0.671	ATP synthase alpha chain (ATP synthesis)
RplO	0.671	ribosomal protein L15
DeoD	0.668	purine nucleoside phosphorylase
OmpF	0.667	major outer membrane protein (porin)
RpsF	0.665	ribosomal protein S6
AckA	0.665	acetate kinase (anaerobic growth: acetate production)
Ppa	0.664	inorganic pyrophosphatase
RpsL	0.662	ribosomal protein S12
LpdA	0.660	dihydrolipoamide dehydrogenase (glycolysis)
AtpD	0.660	ATP synthase beta chain (ATP synthesis)
AdhE	0.659	alcohol dehydrogenase (anaerobic growth in absence of nitate)
Adk	0.652	adenylate kinase
FabB	0.646	3-oxoacyl ACP synthase I (lipids synthesis)
Ndk	0.646	nucleoside diphosphate kinase
RplS	0.643	ribosomal protein L19
Tsx	0.643	nucleoside-specific channel-forming protein (transport)
GuaB	0.640	IMP dehydrogenase (GMP synthesis)
DeoC	0.639	deoxyribose phosphate aldolase ([deoxy]nucleotide catabolism)
SucD	0.637	succinyl-CoA synthetase alpha-subunit (TCA cycle)
PurA	0.636	adenylosuccinate synthetase (AMP synthesis)
AspA	0.633	aspartase
GlnA	0.628	glutamine synthetase (amino-acid synthesis)
ValS	0.626	valyl tRNA synthetase (translation)
SuhB	0.626	extragenic suppressor (heat shock protein related)
Ssb	0.624	single-strand DNA-binding protein (DNA replication)
AceF	0.624	pyruvate dehydrogenase (glycolysis)
RecA	0.621	SOS response
HlpA	0.619	histone like protein HLP-1 (structure)
HtpG	0.617	chaperone (heat shock protein)
AspS	0.613	aspartyl-tRNA synthetase (translation)
PfkA	0.612	6-phosphofructokinase (glycolysis)
Crr	0.611	PTS enzyme III glc (transport)
RplE	0.605	ribosomal protein L5
InfB	0.603	initiation factor IF2 (translation)
LysS	0.602	lysyl tRNA synthetase (translation)
Eda	0.602	2-keto-3-deoxy-6-phosphogluconate aldolase
RplR	0.599	ribosomal protein L18
LeuS	0.599	leucyl-tRNA synthetase (translation)
GuaA	0.599	GMP synthetase (GMP synthesis)
RpoD	0.598	RNA polymerase sigma-subunit (transcription)
SucC	0.594	succinyl-CoA synthetase beta-subunit (TCA cycle)
PstS	0.594	phosphate-specific transport system (transport)
RpsK	0.594	ribosomal protein S11
DsbA	0.592	required for disulphide bond formation
CadA	0.592	lysine decarboxylase (cadaverine production at low pH)
SecB	0.590	protein export protein (transport)
DeaD	0.588	RNA helicase (ribosome assembly)
RplF	0.585	ribosomal protein L6
RpsE	0.583	ribosomal protein S5
GltX	0.581	glutamyl-tRNA synthetase (translation)
SucB	0.580	dihydrolipoamide succinyltransferase (glycolysis)
IcdE	0.579	?
Hns	0.579	Dhistone like protein HLP-2 (structure)
Frr	0.577	ribosme-releasing factor (translation)
FldA	0.576	flavodoxin (electron transport)
FrdB	0.571	fumarate reductase iron-sulfur subunit (TCA cycle)
GlyS	0.571	glycyl-tRNA synthetase (translation)
NirB	0.570	NADH-dependent nitrite reductase (nitrate assimilation)
FrdA	0.569	fumarate reductase (TCA cycle)
AccC	0.568	biotin carboxylase (fatty acid synthesis)
LamB	0.568	maltose et maltodextrines transport
MetK	0.568	S-adenosylmethionine synthetase (one carbon metabolism)
NmpC	0.567	outer membrane porin of endogenous lambdoid bacteriophage
Pgi	0.566	phosphoglucose isomerase (glycolysis)
NarH	0.566	nitrate reductase (nitrate assimilation, induction by nitrate)
MtlA	0.564	mannitol permease (transport)
PtsG	0.564	PTS glucose-specific enzyme II (transport)
RpsH	0.563	ribosomal protein S8
RpoH	0.562	RNA polymerase sigma-32 subunit (heat shock promoters expression)
Apt	0.562	adenine phosphoribosyl-transferase (purine salvage)
CarB	0.562	carbamoyl-phosphate synthetase (Arg & pyrimidine synthesis)
SodB	0.561	superoxide dismutase (radical destruction)
GyrB	0.561	DNA gyrase (DNA replication)
RplQ	0.561	ribosomal protein L17
CydB	0.558	cytochrome D terminal oxidase (electron transport)
Prs	0.558	phosphoribosylpyrophosphate synthetase (nucleotide biosynthesis)

The codon usage of the corresponding genes is good, so that their expressivity level is expected to be high. For instance, the genes for ribosomal proteins, major outer membrane proteins or basic metabolism such as glycolysis belong to this class. Note that genes that are only turned on under special environmental conditions but are abundantly expressed under those circumstances are also present in this class (*e.g.* AckA and AdhE in anaerobiosis, SubH and RpoH after heat shock, NarH in presence of nitrate).

The concentrations of the major tRNA for Lys, Leu and Arg did not follow the general trend. The concentration of the major tRNA for Lys was less than expected and the concentrations of the major tRNAs for Arg and Leu were higher than expected. The concentrations of the major tRNA for Leu and Arg may appear high because their intracellular concentrations do not correspond to their effective availability to the ribosome. For instance, two minor leucyl-tRNAs species are the ones most bound to ribosomes during exponential growth in minimal medium (33). The difference between the effective and measured tRNA concentration could be attributed to the participation of the major leucyl-tRNA species in a reaction other than translation, such as the addition of leucine directly to the amino termini of certain ribosomal proteins (34). This would explain why the effective concentrations of the major tRNA for Leu and Arg could be overestimated from their intracellular concentrations, but does not explain the case of the major tRNA for Lys. However, The comparison of tRNA concentrations from differents authors (35, 36) introduces a note of caution with respect to the interpretation of quantities of tRNA in cells.

Table 5. First 10% of CAI distribution

Prot.	CAI	FUNCTION
RfaL	0.151	O Antigen ligase (LPS core synthesis)
AppY	0.169	transcriptional regulator
RfaS	0.183	LPS core synthesis
RfaK	0.186	LPS core synthesis
TrkG	0.188	integral membrane protein involved in potassium uptake
TdcR	0.189	positive regulatory protein of the tdc operon
PgpA	0.189	Membrane-Bound phosphatidyl glycerophosphate phosphatase
McrC	0.192	Modifies the specificity of McrB restriction
MvrC	0.194	resistance against methyl viologen toxicity
McrA	0.197	methyl cytosine restriction enzyme
PhnQ	0.197	hypothetical protein
Lit	0.198	blocks bacteriophage T4 late gene expression
DacB	0.203	D-alanyl-D-alanine carboxypeptidase in murein metabolism (PBP4)
DsdC	0.204	transcription activator
FimB	0.205	type 1 fimbriae regulatory protein
ThdF	0.207	thiophene oxydation
FimZ	0.210	regulatory protein
FecE	0.211	citrate dependant Fe3+ transport
DicA	0.216	repressor of division inhibition gene dicB
CynR	0.217	transcriptional activator for the cyn operon
RcsA	0.226	transcriptional activator of capsular polysaccharide synthesis
AraJ	0.233	transport or processing of arabinose polymers
BglG	0.233	positive regulator of bgl operon
RfaZ	0.235	LPS core synthesis
Pin	0.236	DNA-invertase
FucU	0.237	?
HemD	0.237	uroporphyrinogen III cosynthetase
PriB	0.238	primosomal replication protein
RnpA	0.240	protein component of ribonucleases P
PgpB	0.241	phosphatidylglycerophosphate phosphatase B
BarA	0.241	OmpR activator
HipB	0.242	?
SulA	0.242	UV-inducible cell division inhibitor
RfaI	0.243	LPS core synthesis
CysX	0.245	hypothetical protein
FimE	0.245	type 1 fimbriae regulatory protein
UmuC	0.246	UV repair enzyme
Iap	0.246	conversion of alkaline phosphatase isozyme
Cdh	0.248	CDP-diglyceride hydrolase
AvtA	0.249	alanine-valine transaminase
RfaP	0.250	LPS core synthesis
HsdS	0.250	EcoE type I restriction-modification enzyme S subunit
BtuC	0.251	cytoplasmic membrane protein involved in vitamin B12 transport
CreB	0.252	transcriptionnal regulatory protein
LysR	0.253	activation of lysA transcription
ProV	0.253	transport of glycine betaine/L-proline
RhaS	0.254	positive activator of genes required for L-rhamnose utilization
FimH	0.254	regulation of length and mediation of adhesion of type 1 fimbriae
DicB	0.254	inhibition of cell division
RfaJ	0.256	LPS core synthesis
UbiC	0.256	chorismate lyase (ubiquinone synthesis)
TdcA	0.257	transcriptional activator for tdc operon
EnvY	0.258	porin thermoregulatory protein
SrlM	0.259	positive regulator for glucitol operon
KgtP	0.260	alpha-ketoglutarate transport
AppA	0.260	acid phosphatase
MiaA	0.260	(delta)2-isopentenyl pyrophosphate tRNA transferase
CadC	0.260	transcriptional activator
MalI	0.261	repressor protein for maltose regulon
OmpT	0.261	outer membrane protease
BioC	0.262	involved in biotin synthesis pathway
LacA	0.263	thiogalactoside transacetylase
McrB	0.263	sequence-specific restriction of cytosine-modified DNA
NlpA	0.263	cytopplasmic membrane liprotein
BicB	0.264	hypothetical protein
RfaY	0.264	LPS core synthesis
GlpR	0.265	repressor of glycerol 3 phosphate regulon
BtuD	0.266	peripheral membrane component of vitamin B12 transport system
FepD	0.266	ferric enterobactin transport protein
RfaB	0.266	LPS core synthesis protein
CybB	0.267	cytochrome b561
HyaF	0.267	protein of hydrogenase-1 operon
AroL	0.268	shikimate kinase II
FliS	0.268	flagellar protein
Pcm	0.268	L-isoaspartyl protein carboxyl methyltransferase type II
PhoQ	0.268	regulation of acid phosphatase
Ogt	0.269	O-6-alkylguanine-DNA-alkyltransferase
Trg	0.269	sensory transducer protein
CreD	0.269	?
AraC	0.270	regulatory protein
CreC	0.270	regulation of CreB
Fes	0.272	enterochelin esterase
Tdk	0.274	thymidine kinase
NlpC	0.274	lipoprotein
SoxR	0.275	regulatory protein for superoxide strength response
MotB	0.277	control of chemotaxis
Dgt	0.277	dGTPase
TyrP	0.278	transport of Tyr
CysC	0.278	APS kinase
Rnc	0.278	ribonuclease III
MetR	0.279	regulatory protein
HyaD	0.279	protein of hydrogenase-1 operon
MutH	0.279	DNA mismatch repair
CysB	0.280	regulatory protein
EntD	0.280	enterobactin synthesis
BolA	0.280	control of cell morphology
BirA	0.281	biotin operon-repressor and biotin holoenzyme synthetase
Phr	0.282	deoxyribopyrimidine photolyase
AroE	0.282	shikimate dehydrogenase
RecF	0.282	control of recombination

The codon usage of the corresponding genes is poor, so that their expressivity level is expected to be low. For instance, many regulatory genes belong to this class.

To validate the interpretation of factor 2, the first and last 10% of the CAI distribution were extracted (Table 4 and 5), and the mean major tRNA frequencies for the proteins were computed in these two extreme classes. The distributions for the two classes were different (figure 8), showing that proteins with high CAI values are enriched in amino-acids carried by the most abundant major tRNA.

Further discussion about the bias in the amino-acid composition of proteins encoded by highly expressed genes should be taken with care because they are based on a logical construction and cannot be directly challenged by experiment. At first glance it seems that it is simpler for tRNAs to adapt their concentration to the amino-acid content of proteins than the reverse because the mutation expense is lower; changing the tRNA concentrations

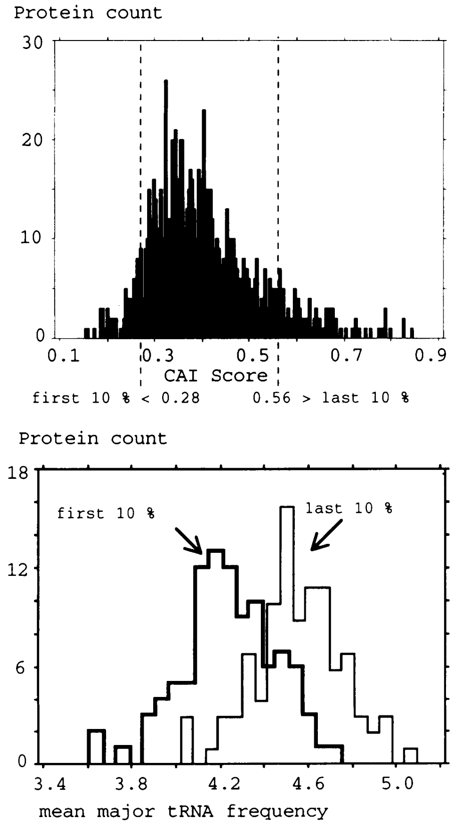

Figure 8. Top: distibution of CAI values for the 999 protein genes of the dataset. The dotted lines indicate the first and last 10% of the distibution. Bottom: distribution of the mean major tRNA frequency for the proteins of the first and last 10% of the CAI distribution.

requires fewer mutation events, such as gene duplication or altered promoter efficiency, than does altering coding sequences, where many sites must be modified. But this cannot explain why the amino-acid compositions of the product of highly expressed genes should be different. This requires that the amino-acid composition of highly expressed genes is particular for some other reason. The simplest explanation is a straightforward adaptation of what is visible at the codon level: highly expressed genes reduce the diversity of codon choices to increase translation efficiency (4). By analogy, proteins encoded by highly expressed genes use a reduced diversity of amino-acid choices to increase translation efficiency.

The fact that proteins encoded by highly expressed genes have a bias of amino-acid usage is an interesting example of the interdependence between translational constraints and overall properties of the protein. The translational constraints seem to be greater than expected since, in addition to selecting the codon corresponding to the most frequent isoacceptor tRNA, they are sufficient to modify the global amino-acid composition. The translational constraints which were known to affect the 'genotype' of proteins, are sufficient to affect their 'phenotype'.

Factor 3 showed that aromatic amino-acids represent a group of amino-acids which frequency is highly variable among proteins. An interpretation is that the biosynthesis of these amino-acids is expensive for the cell, so that there is a selective pressure to reduce the aromaticity of proteins. The fact that these amino-acids are rare (Table 1) is consistent with this hypothesis. However, these amino-acids do not completly disappear, so that there should be an inverse tendency to maintain them in proteins. This inverse tendency could be attributed either to a simple mutationnal drift or more likely to a selective advantage due to a contribution to the stabilization of the three-dimensional structure of the protein.

ACKNOWLEDGEMENTS

We thank M.Gouy, G.Perrière, F.Chevenet and N.Galtier for helpful comments while preparing the manuscript.

REFERENCES

1. Grantham,R. and Gautier,C. (1980) *Naturwissenschaften*, 67, 93–94.
2. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pavé,A. (1980) *Nucl. Acids Res.*, 8, r49–r62.
3. Grantham,R., Gautier,C. and Gouy,M. (1980) *Nucl. Acids. Res.*, 8, 1893–1912.
4. Anderson,S.G.E. and Kurland,C.G. (1990) *Microbiol. Rev.*, 54, 198–210.
5. Gouy,M. and Gautier,C. (1982) *Nucl. Acids Res.*, 10, 7055–7073.
6. Rudd,K.E. (1993) *ASM News*, 7, 335–341.
7. Perrière,G. (1992) PhD thesis #168.92, Lyon I University, France.
8. Gouy,M., Milleret,F., Mugnier,C., Jacobzone,M. and Gautier,C. (1984) *Nucl. Acids Res.*, 12, 121–127.
9. Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and Di Paola,G. (1985) *Comput. Appl. Biosci.*, 3, 167–172.
10. Gouy,M., Gautier,C. and Milleret,F. (1985) *Biochimie*, 67, 433–436.
11. Benson,D., Lipman,D.J. and Ostell,J. (1993) *Nucl. Acids Res.*, 21, 963–2965.
12. Hall,B.G. and Sharp,P.M. (1992) *Mol. Biol. Evol.*, 9, 654–665.
13. Nishikawa,K., Kubota,Y. and Ooi,T. (1983) *J. Biochem.*, 94, 981–995.
14. Miller,C.G. (1987) In Neidhardt,F.C.(ed.) *Escherichia coli and Salmonella typhimurium, cellular and molecular biology*, American Society for Micobiology, Washington, pp. 680–691
15. Sawers,G., Heider,J., Zehelein,E. and BTMck,A. (1991) *J. Bact.*, 173, 4983–4993.
16. Hill, M.O. (1974) *Appl. Statis.*, 23, 340–353.
17. Lebart,L., Morineau,A. and Warwick,K.A. (1984) Multivariate descriptive statistical analysis, John Wiley and Sons, New York.
18. Aitchison,J. (1983) *Biometrika*, 70, 57–65.
19. Thioulouse,J. (1989) *Comput. Appl. Biosci.*, 5, 287–292.
20. Thioulouse,J. (1990) *Comput. Geosci.*, 16, 1235–1240
21. Murtagh,F. and Heck,A. (1987) Multivariate data analysis, Kluwer, Dorchecht, Boston.
22. Perrière,G., Chevenet,F., Dorkeld,F., Vermat,T. and Gautier,C. (1994) *Proceedings of the 27th Hawaii International Conference on System Science* pp 89–97, ACM and IEEE, HI.
23. Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, 157, 105–132.
24. Sharp,P.M. and Li,W.H. (1987) *Nucl. Acids Res.*, 15, 1281–1295.
25. Nishikawa,K. and Ooi,T. (1982) *J. Biochem.*, 91, 1821–1824.
26. Neidhardt,F.C. (1987) In Neidhardt,F.C.(ed.) *Escherichia coli and Salmonella typhimurium, cellular and molecular biology*, American Society for Micobiology, Washington, pp. 3–6.
27. Rogers,S.D., Bhave,M.R., Mercer,J.F.B., Camakaris,J. and Lee,B.T.O. (1991) *J. Bact.*, 173, 6742–6748.
28. Klein,P., Kanahisa,M. and DeLisi,C. (1985) *Biochem. Biophys. Acta*, 815, 468–476.
29. Yamao,F., Andachi,Y., Muto,A., Ikemura,T. and Osawa,S. (1991) *Nucl. Acids Res.*, 19, 6119–6122.
30. Gautier,C. (1987) PhD thesis #87–09, Lyon I University, France.
31. Gouy,M. (1987) PhD thesis #87–52, Lyon I University, France.
32. Shpaer, E.G. (1989) *Protein Seq. Data Anal.*, 2, 107–110.
33. Holmes,W.M., Goldman,E., Miner,T.A. and Hatfield,G.W. (1977) *Proc. Natl. Acad. Sci. USA*, 74, 1393–1397.
34. Goldman,E., Holmes,W.M. and Hatfield,G.W. (1979) *J. Mol. Biol.*, 129, 567–585.
35. Jakubowski,H. and Goldman,E. (1984) *J. Bact.*, 158, 769–776.
36. Ikemura,T. (1981) *J. Mol. Biol.*, 146, 1–21.